Une nouvelle approche non-linéaire pour la segmentation phonétique

Vahid KHANAGHA, Oriol PONT, Khalid DAOUDI, Hussein YAHIA
INRIA Bordeaux Sud-Ouest (Èquipe GEOSTAT)
351 Cours de la Libération, 33405 Talence cedex, France
{vahid.khanagha,oriol.pont,khalid.Daoudi,Hussein.yahia}@inria.fr

Résumé – Le potentiel du Formalisme Multiéchelles Microcanonique (FMM) dans l'identification des frontières de transition du signal de parole a déjà été démontré dans nos travaux antérieurs, en développant une méthode originale de segmentation phonétique. Le FMM repose sur une évaluation précise des exposants de singularité (EdS). Dans ce papier, après avoir décrit en détail un algorithme pour l'estimation précise des EdS dans le cas d'un signal 1D, nous introduisons une nouvelle méthode qui utilise mieux les EdS pour améliorer la précision de la segmentation: d'abord le signal original et une version filtrée sont utilisés pour déterminer un ensemble de frontières candidates; ensuite un test d'hypothèse est effectué sur la distribution des EdS du signal d'origine pour sélectionner les frontières définitives. Nous évaluons la performance de ce nouvel algorithme sur la base TIMIT. Les résultats montrent qu'une amélioration considérable des performances de segmentation est réalisée.

Abstract – The potential of the Microcanonical Multiscale Formalism (MMF) in the identification of transition fronts in speech signal was demonstrated in our earlier work by developing an original phonetic segmentation method. The MMF relies on the calculation of Singularity Exponents (SE). In this paper, after describing the detailed algorithm for precise estimation of SE in the case of a 1D signal, we introduce a novel method of segmentation which aims at further exploiting the capability of SE in phoneme boundary identification: first speech signal and its low-passed version are used to detect a set of candidate boundaries and then a hypothesis test is performed over the distribution of SE of the original signal to select the final boundaries. We evaluate our algorithm on the TIMIT database. The results show that a considerable improvement in segmentation performance is achieved.

1 Introduction

Le caractère turbulent et non-linéaire du signal de parole est bien établi [1, 2], cependant la tendance dominante en traitement de la parole est basée sur des approches linéaires (notamment à travers le modèle source-filtre). En considérant le signal parole comme l'acquisition d'un système complexe au sens physique, nous introduisons un cadre radicalement nouveau pour l'analyse non linéaire du signal de la parole. Notre approche est basée sur le Formalisme Multiéchelles Microcanonique (FMM) qui repose sur des concepts et principes provenant de la physique statistique des systèmes complexes et turbulents. Le FMM [3] est basé sur le calcul précis des Exposants de Singularité (EdS) dont la distribution contient des informations clés sur la dynamique intermittente du signal. Dans ce papier, nous commençons par décrire en détails un algorithme pour l'estimation précise des EdS. Dans [4], nous avons montré que ces derniers véhiculent des informations sur la dynamique locale de la parole et qui peuvent être facilement utilisées pour détecter les frontières entre phonèmes. Nous avons ensuite proposé une méthode automatique et efficace pour la segmentation phonétique indépendante du texte.

Dans cet article, en faisant une analyse d'erreurs de notre algorithme original, nous proposons une technique en 2 étapes qui exploite mieux les EdS pour améliorer la performance de la segmentation. La première étape consiste à utiliser notre algorithme original pour détecter les frontières candidates, sur le signal et sur une version filtrée passe-bas. Dans la deuxième étape, nous utilisons un test d'hypothèse sur la distribution locale des EdS pour sélectionner les frontières qui correspondent à un véritable changement de distribution. Nous évaluons la performance de ce nouvel algorithme sur toute la partie Train de la base TIMIT [5] et nous la comparons avec une technique récente en état-de-l'art [6]. Les résultats montrent qu'une amélioration considérable des performances est réalisée.

Cet article est organisé de la façon suivante : nous introduisons les EdS en section 2 puis nous détaillons leur calcul en section 3. La section 4 présente brièvement notre algorithme précédent de segmentation. La nouvelle méthode est décrite dans la section 5, tandis que les résultats expérimentaux sont présentés dans la section 6.

2 Les exposants de singularité

Des approches récentes en analyse des signaux complexes, et tout particulièrement en analyse du signal Parole consistent à considérer un signal comme une acquisition d'un système dynamique complexe [7]. Dans ce cadre, certaines quantités associées à la prédictabilité dans ces systèmes peuvent être déterminées sur les signaux d'acquisition, par exemple les exposants de Lyapunov. Dans ce travail, nous nous intéressons aux EdS, calculés dans le cadre du FMM, qui ont prouvé leur puissance dans le cadre du formalisme des systèmes reconstructibles : ils ont été utilisés dans une grande variété d'applications allant de la compression de données à l'inférence et la

prédiction [8]. Ces exposants sont définis par l'examen d'une loi de puissance multi-échelles [3]. Étant donné un signal s, pour au moins une fonctionnelle Γ_r dépendante de l'échelle r, la relation suivante doit être valide à tout instant t :

$$\Gamma_r\left(s(t)\right) \;=\; \alpha(t)\, r^{h(t)} + \mathrm{o}\left(r^{h(t)}\right) \qquad r \to 0 \qquad (1)$$

où $h(t)$ est l'EdS du signal s à l'instant t. Les EdS quantifient le degré de prédictabilité : plus h(t) est petit moins le système est prédictible en t. Turiel et al. [3] ont proposé un choix pour la fonctionnelle Γ_r défini à partir des caractérisations typiques de l'intermittence en turbulence :

$$\Gamma_r\left(s(t)\right) \;:=\; \frac{1}{r}\int_{]t-r,t+r[} \mathrm{d}\tau\; |s'(\tau)| \qquad (2)$$

où s' est le gradient de s. Le problème du calcul des EdS est détaillé dans la section suivante.

3 Calcul des exposants de singularité

La fonctionnelle 2, définie à partir du gradient du signal, peut être projetée en ondelettes de manière à obtenir des interpolations continues à partir de données discrètes échantillonnées : si Ψ est une ondelette, la projection de la fonctionnelle Γ_r au point t et à l'échelle r est définie par l'équation [9] :

$$\mathcal{T}_\Psi \Gamma_r(t) = \frac{1}{r}\int_{-r}^{r} |s'(\tau)| \Psi\left(\frac{t-\tau}{r}\right) \mathrm{d}\tau \qquad (3)$$

Si le signal s vérifie l'équation 1, alors la projection de la mesure du gradient de s selon l'équation 3 vérifie une équation similaire avec le même EdS $h(t)$ [10] ce qui conduit à une estimation simple des EdS par régression $\log - \log$ sur une projection en ondelettes en chaque point t [11]. Le pouvoir de résolution d'une ondelette dépend du nombre de passages par zéros de son graphe qui est donc minimal pour les ondelettes positives. En conséquence l'introduction de la mesure basée sur le gradient (2) améliore la résolution spatiale nécéssaire à l'estimation des exposants de singularité.

Cependant, il est possible d'aller encore plus loin dans la précision de l'estimation lors du calcul des EdS, notamment lorsque l'on cherche à atténuer les phénomènes d'oscillations propres à une décomposition en ondelettes, ou bien lorsque l'on veut éviter le problème de la détermination d'une ondelette adaptée à la nature du signal. Dans [12] est présenté un algorithme d'évaluation des EdS basé sur l'évaluation d'une mesure associée à la variété des points non-prédictibles UPM (Unpredictable Points Manifold) qui quantifie le degré de reconstruction locale dans un signal, c'est à dire le degré de prédictibilité en un point, quantité directement reliée à l'EdS en ce point. Dans la suite de cette section, nous détaillons cette méthode de calcul des EdS dans le cas du signal Parole.

La remarque fondamentale consiste à observer que la formule de reconstruction associée à la variété la plus singulière (définie par l'ensemble des points du signal dont l'EdS est le plus petit à une certaine précision numérique), telle que celle exposée dans [8] utilise un noyau de reconstruction qui prend la forme de l'inverse d'un gradient dans l'espace Fourier :

$$g(\hat{\mathbf{k}}) = i\frac{\mathbf{k}}{|\mathbf{k}|^2} \qquad (4)$$

Les points du signal qui conduisent à une reconstruction parfaite sont les moins prédictibles, c'est à dire que la valeur du signal en un tel point ne peut être déduite de celle de ses voisins. Nous allons par conséquent définir une quantité associée au degré de prédictibilité local en chaque point. Cette quantité est une mesure vectorielle spéciale définie par une projection en ondelettes du gradient qui pénalise la imprédictibilité. Cette mesure vectorielle définit à son tour un ensemble géométrique, dans le domaine du signal, la variété des points les moins prédictibles, et l'hypothèse fondamentale posée dans cette méthode de calcul des EdS est que la variété des points les moins prédictibles est identique à la variété la plus singulière, c'est à dire l'ensemble des points du signal ayant l'exposant de singularité minimal (à une précision donnée) [12]. Étant donné un point t du domaine d'un signal discret s, le voisinage le plus simple associé à la prédictibilité en t est formé des trois points (p_0, p_1, p_2) avec $p_0 = t$, $p_1 = t+1$ et $p_2 = t-1$. Pour éviter l'utilisation des harmoniques standards $(e^{2ik\pi/n})_k$, qui dépendent de la taille n, on remarque que la fréquence de Nyquist la plus simple dans les deux directions autour de t d'un signal 1D est $2\pi/3$. Par conséquent nous introduisons le nombre complexe $j = e^{2i\pi/3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$, $\bar{j} = j^2$ ainsi que l'opérateur de matrice :

$$\mathcal{F} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & j & \bar{j} \\ 1 & \bar{j} & j \end{bmatrix} \qquad (5)$$

Définissons maintenant, dans l'espace Fourier, un opérateur de dérivation naturellement associé à une demi-différence entre un point et ses voisins immédiats.

Puisque $\sqrt{3} = 2\sin(\pi/3)$ nous définissons l'opérateur $\hat{d_x} = (0, i\sqrt{3}, -i\sqrt{3})$ dont l'action sur un vecteur se fait selon chaque composante puis $d_x = \mathcal{F}^{-1}\hat{d_x}\mathcal{F}$ (i.e. on multiplie $\mathcal{F}$ par un vecteur puis on applique $\hat{d_x}$ et on multiplie le vecteur obtenu par $\mathcal{F}^{-1}$) ainsi que l'opérateur de reconstruction local $\mathcal{R} = \mathcal{F}^{-1}\hat{\mathcal{R}}\mathcal{F}$ avec $\hat{\mathcal{R}} = (0, -i\sqrt{3}, i\sqrt{3})$. Ces opérateurs de gradient et de reconstruction nous servent à définir une mesure UPM de corrélation locale de la façon suivante. Étant donnés un point t_0 et une échelle r_0, on définit le voisinage de t_0 comme ci-dessus (p_0, p_1, p_2) et le vecteur signal associé à ce voisinage (s_0, s_1, s_2). Etant donnée la moyenne $\bar{s} = \frac{1}{3}(s_0 + s_1 + s_2)$ on forme le vecteur "redressé" (u_0, u_1, u_2) avec $u_0 = p_0 + \bar{s}$, $u_1 = p_0 - \bar{s}$, $u_2 = p_0 - \bar{s}$. On applique l'opérateur d_x au vecteur (u_0, u_1, u_2) pour obtenir le vecteur (g_0, g_1, g_2) dont on sauvegarde la première composante $A = g_0$. L'opérateur de reconstruction locale est ensuite appliqué à (g_0, g_1, g_2) pour en déduire un signal reconstruit (q_0, q_1, q_2), auquel on ré-applique l'opérateur de gradient d_x pour obtenir (ρ_0, ρ_1, ρ_2). La mesure UPM de corrélation locale est alors définie par l'égalité :

$$\mathcal{T}_{\Psi_{lcsm}} \Gamma_{r_0}(t_0) = |A - \rho_0| \qquad (6)$$

d'où l'on déduit un EdS selon une procédure identique à celle exposée dans [12].

4 Application à la segmentation des signaux de parole

Dans [4], nous avons d'abord montré que la distribution conditionnelle des EdS présente des changements clairs à la frontière des phonèmes. En exploitant l'interprétation la plus simple de ces changements, la variation des moyennes, nous avons défini ensuite une mesure sur les EdS pour mieux mettre en évidence ces changements :

$$ACC(t) = \int_{t_0}^{t} \mathrm{d}\tau \; h(\tau) \tag{7}$$

La figure 1 montre un exemple d'un signal de parole (de la base TIMIT) et de la fonction ACC associée, les lignes verticales représentant la segmentation phonétique manuelle donnée dans TIMIT. On peut voir qu'à l'intérieur de chaque phonème ACC est presque linéaire et qu'il y a des changements abrupts de pente aux frontières des phonèmes. Pour le développement d'un algorithme automatique de segmentation, nous avons ajusté une courbe linéaire par morceaux à ACC et identifié ses points de rupture comme étant les frontières entre phonèmes.

5 Amélioration de l'algorithme de segmentation

Dans ce papier nous présentons une technique qui exploite encore mieux le potentiel des EdS pour améliorer la précision de la segmentation. En procédant à une analyse d'erreurs de notre algorithme, nous avons observé que certaines frontières manquées correspondent à des transitions entre les fricatives-/stops et les voyelles. Nous avons également observé que les transitions entre la parole active et des segments à faible énergie (tels que les pauses et le silence épenthétique) correspondent à des changements de pente nets dans ACC et sont ainsi faciles à détecter. En effet, les EdS sur les segments de basse énergie ont des valeurs positives élevées, alors qu'ils ont souvent des valeurs négatives dans les segments de parole active. Ces observations et le fait que les fricatives/stops sont essentiellement des signaux à haute bande, nous ont motivé pour calculer ACC sur une version filtrée passe-bas du signal. Ce faisant, ces transitions seront converties en transitions silence-parole qui sont beaucoup plus faciles à détecter. Comme il est connu que l'essentiel de l'énergie spectrale des fricatives est situé au-dessus de 2000Hz, et que pour la plupart des stops les bandes de fréquences actives commencent à 1800Hz, nous avons donc choisi la fréquence de coupure du filtre passe-bas à 1800Hz.

D'autre part, nous avons observé que certaines frontières manquées correspondent à des phonèmes voisins qui présentent une différence assez distinctive dans leur distribution EdS ; toutefois, le changement de leur moyenne n'est pas assez fort pour être traduit par un changement dans la pente de ACC (ainsi il n'est pas capturé par la procédure simple d'ajustement de courbe). Cela nous conduit à utiliser un test d'hypothèse statistique sur les distributions des EdS afin de détecter de telles frontières.

Notre nouvel algorithme de segmentation est donc le suivant. Dans la première étape, nous utilisons notre algorithme original pour détecter les frontières dans le signal original et sa version filtrée. Nous recueillons toutes les frontières détectées et les considérons comme des candidats. Dans la deuxième étape, nous prenons la décision finale en effectuant un fenêtrage dynamique sur ces candidats suivi d'un Test de Rapport de Vraisemblance (TRV) sur les distributions des EdS du signal original. Nous utilisons une hypothèse Gaussienne car notre but est de détecter des changements dans la moyenne et la variance. Plus précisément, pour chaque candidat c_i nous considérons la grande fenêtre $Z = [c_{i-1}, c_{i+1}]$ et les deux petites fenêtres $X = [c_{i-1}, c_i]$ et $Y = [c_i, c_{i+1}]$. Nous calculons ensuite la statistique du TRV pour décider entre les deux hypothèses :

- H_0 : les EdS de Z sont générés par une seule gaussienne.
- H_1 : les EdS de Z sont générés par deux gaussiennes sur X et Y.

Si H_1 est choisie nous sélectionnons c_i comme étant une frontière, sinon, c_i est retiré de la liste des candidats. Nous soulignons ici que les EdS du signal filtré ne sont utilisés que dans la première étape. La décision finale est faite sur l'information portée par les EdS du signal original. Nous soulignons également que ce nouvel algorithme est toujours aussi simple et rapide que celui d'origine.

6 Résultats expérimentaux

L'évaluation est effectuée sur toute de la partie Train de la base de données TIMIT [5] qui contient 4620 phrases prononcées par 462 locuteurs. Nous comparons les performances de notre nouvel algorithme avec l'original et aussi une méthode état de l'art [6]. Les résultats pour une fenêtre de tolérance de 15ms sont reportés dans le tableau 1, en terme de la mesure de performance F_1 [13], et aussi d'une mesure proposée récemment appelée $R - value$ [14].

Les résultats montrent qu'une amélioration de 2% est atteint par rapport à notre méthode originale d'utilisation des EdS (ACC) pour la tolérance de 15ms. En outre, comparativement à la référence [6], une amélioration de l'ordre de 10% est atteint. Ces résultats prouvent la puissance des EdS dans la localisation précise des frontières de phonèmes.

7 Conclusion

En adaptant au cas des signaux 1D un algorithme précis de calcul des EdS nous introduisons un nouvel algorithme de segmentation qui non seulement surpasse notre méthode originale,

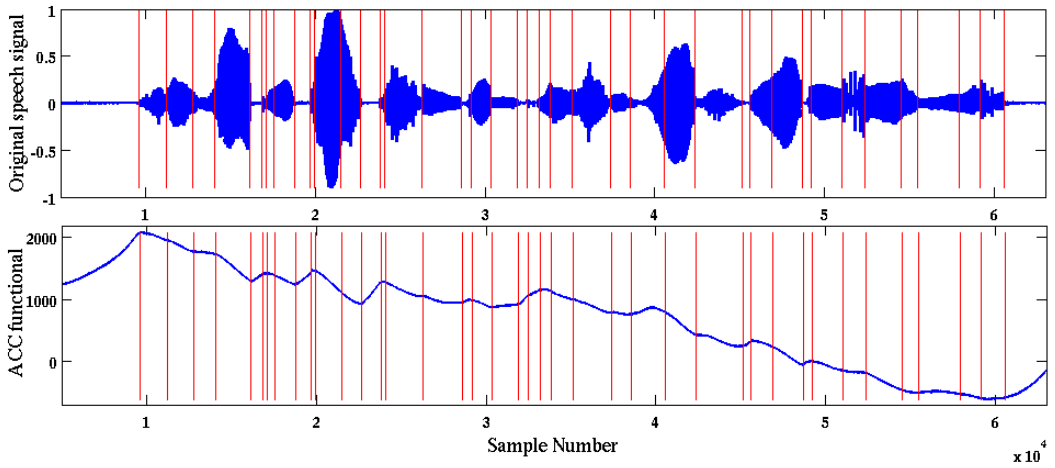

FIG. 1 – **HAUT :** Un signal de parole. **BAS :** La fonctionnelle proposée pour l'identification des frontières de phonèmes.

TAB. 1 – Comparaison des performances pour une fenêtre de tolérance de $= 15ms$

	Dusan et al [6]	ACC [4]	ACC+TRV
F_1	0.55	0.63	0.65
$R - value$	0.60	0.68	0.70

mais est aussi nettement plus précis que les algorithmes de l'état de l'art. Nous avons ainsi montré que les EdS constituent un outil puissant et prometteur pour la segmentation de la parole. Au delà de la segmentation, ces résultats encourageants (obtenus par une approche radicalement nouvelle en parole) suggèrent que le FMM a un grand potentiel en analyse de la parole et mérite d'être étudié plus profondément.

Références

[1] G. Kubin, *Nonlinear processing of speech. Chapter 16 on Speech coding and synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995.

[2] S. McLaughlin and P. Maragos, *Nonlinear methods for speech analysis and synthesi, in Advances in Nonlinear Signal and Image Processing*, S. Marshall, E. B. S. o. S. P. G. L. Sicuranza, and Communications, Eds. Hindawi Publ. Corp., 2006.

[3] A. Turiel, H. Yahia, and C. P. Vicente., "Microcanonical multifractal formalism : a geometrical approach to multifractal systems. part 1 : singularity analysis," *J. Phys. A, Math. Theor.*, vol. 41, p. 015501, 2008.

[4] V. Khanagha, K. Daoudi, O. Pont, and H. Yahia, "A novel text-independent phonetic segmentation algorithm based on the microcanonical multiscale formalism," *Proceedings of INTERSPEECH*, 2010.

[5] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "DARPA TIMIT acoustic-phonetic continuous speech corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, Tech. Rep., 1993.

[6] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," *Proceddings of INTERSPEECH/ICSLP 2006*, pp. 645–648, 2006.

[7] I. Kokkinos and P. Maragos, "Nonlinear speech analysis using models for chaotic systems," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1098–1109, Jan. 2005.

[8] A. Turiel and A. del Pozo, "Reconstructing images from their most singular fractal manifold," *IEEE Trans. on Im. Proc.*, vol. 11, pp. 345–350, 2002.

[9] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[10] A. Turiel and N. Parga, "The multi-fractal structure of contrast changes in natural images : from sharp edges to textures," *Neural Computation*, vol. 12, pp. 763–793, 2000.

[11] A. Turiel, C. Pérez-Vicente, and J. Grazzini, "Numerical methods for the estimation of multifractal singularity spectra on sampled data : A comparative study," *Journal of Computational Physics, Volume 216, Issue 1, p. 362-390.*, vol. 216, pp. 362–390, 2006.

[12] O. Pont, A. Turiel, and H. Yahia, "An optimized algorithm for the evaluation of local singularity exponents in digital signals," in *14th International Workshop on Combinatorial Image Analysis*, 2011.

[13] A. Esposito and G. Aversano, "Text independent methods for speech segmentation," in *Nonlinear Speech modelling. G. Chollet et al Eds.*, 2004, pp. 261–290.

[14] O. J. Rasanen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure : the R-value," *Proceedings of INTERSPEECH*, 2009.